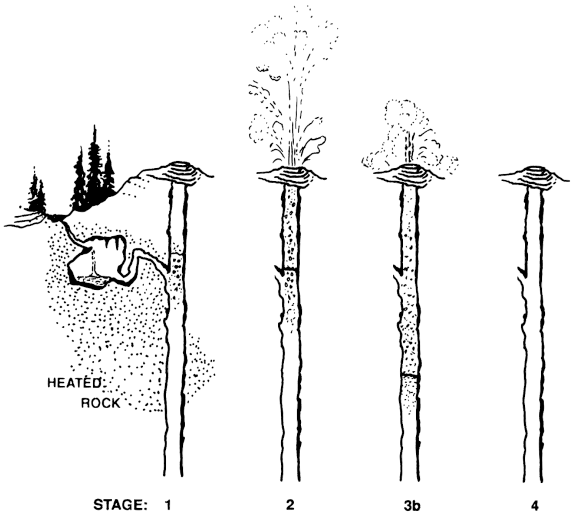
Old Faithful



---

66 marks

In the Yellowstone National Park, Wyoming, USA there is a famous geyser called "Old Faithful" which erupts with some regularity. The physical model is thought to be something like the illustration (from Rinehart (1969), p. 572, via Azzalini and Bowman (1990)):

Azzalini and Bowman (1990) describe what's happening in the stages as follows:

Stage 1. "The tube is full of water which is heated by the surrounding rocks. The water is heated above the normal boiling temperature because of increased pressure. This due to the mass of water which is on top: the deeper the water the higher the temperature required for boiling. Moreover, 'whereas the water in the tube is superheated with respect to the ambient boiling point at the mouth of the geyser, the water temperature at depth is far below the boiling point curve that must be applied to a vertical column of water'. "

Stage 2. "When the top water reaches the boiling temperature, it becomes steam and moves towards the surface. The pressure at the bottom then drops rapidly to the normal level and, by an induction effect, the bottom water rapidly becomes steam. This cascading mechanism is repeated several times: as water is converted into steam, the pressure on lower water is decreased, causing the production of more steam and triggering the eruption."

Stage 3a. "If at the time of cascading the temperature in the lower regions is lower than might be expected, cascading stops short of the bottom and the play is short.' Stage 3b. Alternatively, 'when the temperature is comparatively high at these depths, cascading works itself down much farther and the play is long'."

Stage 4. "The geyser tube is completely or partly empty, ready to be filled with new water."

"We do not discuss geological reasons for the fact that sometimes the cascading effect works down to the bottom of the tube while at other times it stops earlier. We simply note the phenomenon and discuss its consequences. Stages 3a and 3b are associated with short and long waiting times for the next eruption. In stage 3a, the system starts a new cycle partially filled with hot water so that the following heating time is shorter; at the new eruption the entire tube will be emptied, since part of the water had already been heated in the previous cycle."

For each eruption, the waiting time w between its beginning and the beginning of the previous eruption is recorded to the nearest minute and the duration d of the eruption is recorded to fractions of a minute.

Collected from August 1st until August 15th, 1985 the data record the 299 successive eruptions which occurred during this time. Though R. A. Hutchinson, the park geologist, collected similar data sets, it is not clear from the source whether or not this data set is one of them. Measurements had to be taken through the night and duration times for these eruptions were recorded only as being one of short, medium, or long (encoded here as 2, 3, or 4 minutes, respectively).

The data on `duration` and `waiting` times are contained in the `geyser` data set found in the `MASS` package. Load this as

```
library(MASS)
data(geyser)
```

a. *(3 marks)* Describe the target population/process $\mathcal{P}_{Target}$ you think scientific investigators have in mind for the above problem. Carefully define both what constitutes an individual unit of $\mathcal{P}_{Target}$ and how the set of units is defined.

b. *(4 marks)* Describe a study population/process $\mathcal{P}_{Study}$ as it might have been available for the scientific investigators. Again, carefully define both what constitutes an individual unit of $\mathcal{P}_{Study}$ and how the set of units is defined. Why might there be study error?

c. *(4 marks)* Describe the sample $\mathcal{S}$. Again, carefully define both what constitutes an individual unit of $\mathcal{S}$ and how the set of units is defined. Why might there be sample error?

d. *(2 marks)* Imagine the process for selecting a sample. How might this process produce sampling bias?

e. *(4 marks)*
Given the above description of a physical model for how the geyser might work, explain why the independence of the variates in each of the following pairs might be of interest:

   i. w_i and d_i
   ii. d_i and w_{i+1}
   iii. d_{i-1} and d_i
   iv. w_{i-1} and w_i

f. *(2 marks)* Describe one other variate of potential interest which is implicitly defined in this data set? How would you determine its value?

g. *(3 marks)* Imagine the measuring process. What problem(s) do you think might be associated with the measuring process? How might it manifest itself in terms of measuring bias and/or variability?

h. *(10 marks)* To assess the measuring systems, we might consider looking at the least significant parts of each measurement. For this the modulus arithmetic binary operator `%%` in `R` can be handy to find the least significant part of a measurement. For example `x %% 10` will return the rightmost digits in a non-negative integer `x` and `x %% 1` will return the fractional part of a non-negative real number `x`.

   Using the `%%` modulus operator to construct the appropriate data set, perform a Pearson chi-square goodness of fit (in each case use 10 non-overlapping equal size bins) to test each of the following hypotheses:

   i. H_d the fractional part of the duration follows a $U[0,1]$ distribution,
   ii. H_w the rightmost digit of the waiting time equiprobably any one of the digits $0, 1, 2, \ldots, 9$.

   Summarize your findings (including showing your code). What do you conclude about the two measuring systems?

i. *(12 marks)* Plot the sample quantiles of both the duration and the waiting times on the same plot (use a different colour for each variate). Show your plot and the code used to generate it. By referring to the relevant features of the sample quantiles, separately describe the distribution of each variate and compare the two distributions to one another. Now compare the two distributions by constructing an appropriate quantile-quantile plot and referring to its relevant features. Again show the plot and the code.

j. *(10 marks)* Consider the waiting times w_i. We might ask whether waiting times could have been independently distributed. One way to test this is to compare each waiting time w_i with that one that occurred exactly k eruptions previously, namely w_{i-k}, the so called "lagged k" value. For $k \geq 1$, there will be $n - k$ pairs (w_{i-k}, w_i) which could be assessed for independence. A scatterplot of these pairs could be used to assess independence.

Alternatively, we might first transform them to values which should be more nearly uniformly distributed. To that end, define

```
transform2uniform <- function(x,
                              a = if(length(x) <= 10) 3/8 else 1/2,
                              ...) {
  (rank(x, ...) - a) / length(x)
}
```

Now use the function `transform2uniform()` on the waiting times to give values $u_i = \widehat{Q}_W(w_i)$. You will now consider the independence of u_i and its lag k value u_{i-k}. If they are independent, the scatterplot of the $n-k$ pairs (u_{i-k}, u_i) should look like **uniform scatter** in the unit square.

Conduct a scatterplot line up test for independence of u_{i-k} and u_i for each of

i. $k = 1$, the immediately preceding eruption, and
ii. $k = 22$, the eruption occurring roughly the day before.

Show your code for constructing the necessary data and the lineup plots. What do you conclude about the dependence between waiting times?

k. *(12 marks)* Consider the possible dependence of the ith duration d_i on that duration, d_{i-k}, lagged k behind. Using a two-dimensional kernel density estimate as a means to display the data (without the data points), conduct a lineup test of independence using joint density contours for each of

i. $k = 1$, the immediately preceding eruption, and
ii. $k = 22$, the eruption occurring roughly the day before.

Show your code for constructing the necessary data and the lineup plots. What do you conclude about the dependence between durations lengths?